Molecular Biology IV: Microarrays and Bioinformatics

I. Learning Objectives

A. Explain how microarrays are used for:

1. expression profiling.
2. single nucleotide polymorphism (SNP) detection.

B. Describe how bioinformatics integrates biological information and together with molecular methods is leading to the identification of disease susceptibility genes and the development of future risk predictors, diagnostic tests and therapeutic targets.

II. Oligonucleotide microarrays or "chips."

A. **Introduction:** There are still many thousands of genes in every eukaryotic cell that we know nothing about. We also only know a few of the genes that make us susceptible to diseases. To gain insight into the function of genes, gene expression changes are being studied throughout embryonic development, aging, and disease progression and treatment. The above information is being obtained faster than ever by using DNA hybridization microarrays, a new technology only in its initial developmental stages.

B. **Major refinements** of the technology underlying DNA libraries, PCR, and hybridization have come together in the development of DNA microarrays (sometimes called DNA chips). In Southern blotting, the DNAs being tested are immobilized on a membrane, and labeled probe(s) in solution are hybridized with the target DNA (tester DNA) on the membrane to detect the presence of specific genes. In contrast, for microarray methods *thousands of different DNA probes are arrayed in known positions on a solid support and tester DNA is labeled with a fluorescent tag and hybridized with the array*. Hybridization of the fluorescent-tagged tester DNA is detected using a fluorescence reader. The reader's computer knows the probe's DNA nucleotide sequence at each position on the array; fluorescence at a given position indicates that the fluorescent-tagged tester DNA contains the DNA complimentary to the probe attached to the solid surface. Microarray technology allows tens of thousands of nucleic acid sequences to be investigated at a time. It is sensitive enough to detect a single base pair difference between two tester DNAs. For example, it can be used to determine if an individual is heterozygous for a single nucleotide polymorphism (SNP). It can also be used to determine which mRNAs are present in different cell types and under different conditions (gene expression profiling).

C. **Types of microarrays**. Any DNA sequence, from any source, can be used as a solid-surface probe on a microarray. The tester DNA can be cDNA or genomic DNA. Two types of microarrays are typically in use today.

1. ***Oligonucleotide microarrays*:** Oligonucleotides are synthesized directly on the solid support (generally less than 25 nucleotides long). Hybridization conditions are chosen that distinguish single-base mismatches from perfect hybrids.
2. ***DNA fragment microarrays*:** DNA fragments (hundreds of nucleotides long) are stamped onto glass microscope slides using robotic devices that accurately deposit nanoliter quantities of DNA solutions in a pre-designed array onto a surface measuring only a few cm^2. Usually the stamped DNA is made by PCR.

D. **Sources of tester DNA**. Small samples of cells from patients (individuals) are the source of templates to create the tester DNA.

1. Genomic DNA - multiple sets of PCR primers are used to amplify and fluorescently tag different regions of genomic DNA for the detection of mutations or polymorphisms.
2. mRNA – reverse transcription is used to synthesize and fluorescently tag the 1st cDNA strand.

E. **Ways microarrays are used**:

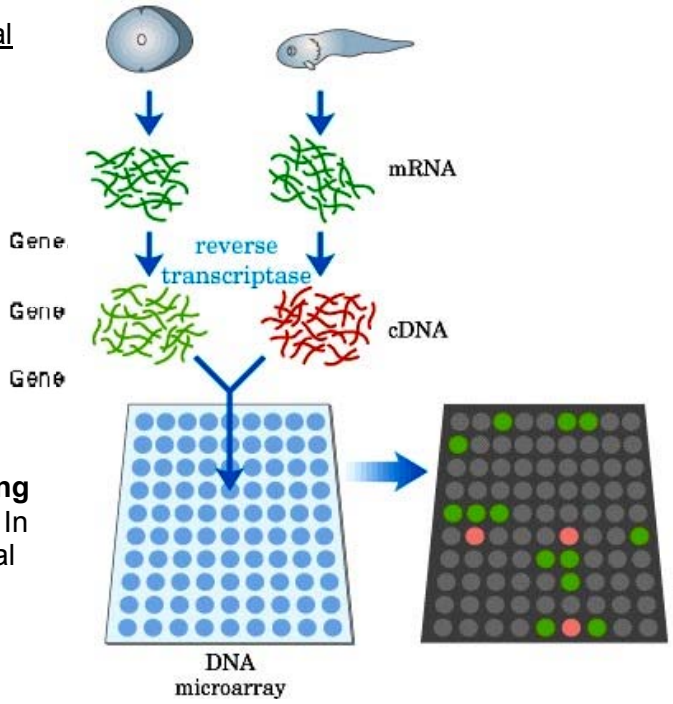
1. *gain of signal* (*a*): A single type of fluorescent signal appears at specific spots on the array. The places where the tester DNA hybridizes indicates the types of tester DNA present.
2. *two color loss-of-signal* (*b*): A fluorescent colored reference standard and a different fluorescent colored tester DNA are co-hybridized with the microarray. The amount of standard DNA that is displaced from a spot can indicate the quantity of the tester DNA relative to the standard DNA.
3. *minisequencing* (*c*): The tester DNA is hybridized to an oligonucleotide array such that the polymorphic nucleotide position serves as the template for addition of the first nucleotide to the 3' end of the solid phase probe. All four dideoxy-nucleotides are added to the chip together with a polymerase. The specific fluorescent dideoxy-nucleotide added to the 3' end of the oligonucleotide on the chip indicates which polymorphism the individual has.

F. **Major uses of microarrays**:

1. Expression profiling - Identifies the repertoire of mRNA made by the cell type being assayed. Both "DNA fragment microarrays" and "oligonucleotide microarrays" can be used for this application.
2. SNP detection – Determines differences in nucleic acid sequence among individuals. "Oligonucleotide microarrays" are typically used for this application.

G. **Gene expression profiling**.

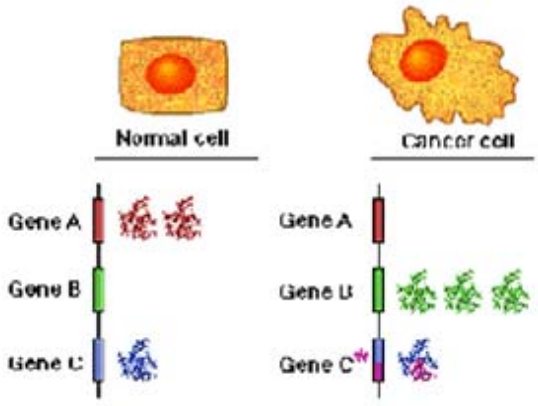
1. Gene Expression distinguishes normal cells. "Although every cell contains the full set of ~30,000 genes, only some of these are expressed in any particular type of cell. Cells look and act the way they do because of the specific genes that they express and the amounts of gene products produced. For example, a muscle cell, a skin cell, and a nerve cell, could be distinguished by their gene expression profiles."

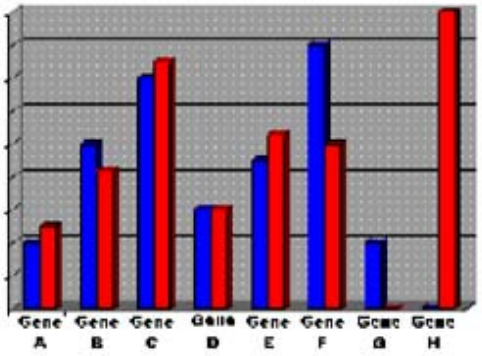


H. **An example of expression profiling using the "two color loss-of-signal" method**. In the example in the adjacent figure, the total

mRNA is isolated from cells in two different stages of development. The mRNA from each cell type is separately converted into fluorescent-tagged cDNA using *reverse transcriptase* to incorporate fluorescent-labeled deoxynucleotides into 1st strand cDNA. One mRNA is converted into fluorescent-**red 1st strand cDNA** and the other mRNA is converted into fluorescent-**green 1st strand cDNA**. The two different cDNA are **pooled and added to a microarray of gene probes**. The fluorescent cDNAs hybridize to complementary sequences on the microarray. After washing away of the fluorescent-tagged 1st strand cDNA that does not hybridize, a fluorescence reader is used to determine the red fluorescent intensity and the green fluorescent intensity at each spot on the array. The fluorescent cDNAs are quantitatively representative of the mRNAs from which they were synthesized and therefore, the fluorescence intensity of a spot measures the relative abundance of that particular mRNA in the cells. In the example in figure (which is not in color), spots that fluoresce green represent mRNA that is more abundant in the cell stage on the left; those that fluoresce red represent sequences more abundant later in development (cell stage on the right). A yellow spot is sometimes used to indicate a similar level of mRNA in both stages. Microarrays can provide a snapshot of all the genes being expressed in cells at the moment they were harvested - providing a measurement of gene expression on genome wide scale. For a gene of unknown function, the time and circumstances of its expression can provide important clues about its role in the cell.

I. **Clinical Correlation: Gene Expression in normal vs. cancer cells**. "The repertoire of gene products produced by a cancer cell might differ in two ways from its normal counterpart: 1) Quantitatively, as shown for gene B, which is expressed at an abnormally high level, and gene A, which is not expressed at all. 2) Qualitatively, as shown for gene C, which is mutated such that it produces an altered protein" (see adjacent figure). **Why the need for molecular classification of tumors?** Historically, tumors have been classified based on their morphological appearance. This classification is of limited utility because tumors with similar histopathological appearance often follow different clinical courses (aggressiveness and response to therapy). Molecular subclassification of tumors using gene expression profiling identifies the genetic differences among normal cells, precancerous cells, and cancer cells. Establishing for a cell the repertoire of genes expressed, together with the amount of gene products produced for each, yields a powerful 'fingerprint'. Comparing the fingerprints of a normal versus a cancer cell identifies genes that by their suspicious absence or presence (such as gene H in the figure below) deserve further scientific scrutiny to determine whether they play a role in cancer, or can be exploited in a test for early detection.

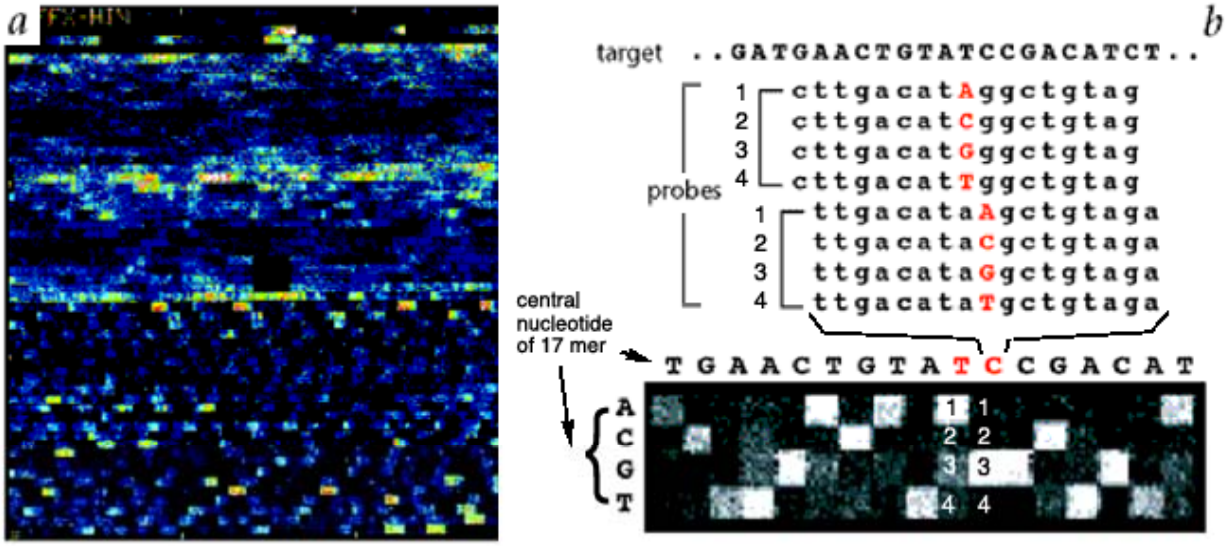




Thus, these methods are improving the ability to detect cancer and to match patients with appropriate treatment.

J. **"Oligonucleotides arrays for identifying SNPs, testing for correlations between SNPs and disease (linkage analysis), and for genetic testing**.

1. Identification of SNPs that are useful for linkage analysis. "Oligonucleotide" arrays can be used to determine the nucleotide sequence of DNA. Genomic DNA of many individuals in a population can be "sequenced" with microarrays in order to identify commonly occuring polymorphisms. Since the human genome sequence



has been determined, the "reference sequence" for this region is known and can used to design an "oligonucleotide array" chip to determine, within a DNA region, the frequency and distribution of nucleotide variation in a population. Genomic DNA is isolated from individuals and then PCR is used to amplify as well as fluorescent-tag it. This tester DNA is hybridized to the "oligonucleotide microarray," which can distinguish single-base mismatches from perfect hybrids. The microarray for this purpose (see figure below) contains a large number of 17mer oligonucleotides on each spot. Each spot contains a 17 mer that is complementary to a region on the tester DNA. A number of spots are present on the array so that a complimentary 17mer is present for each and every 17mer along the linear length of the DNA region being tested. Additionally, the central position of each 17 mer probe is varied to be A, T, C or G; i.e., four 17 mer probes, each on a different spot, are present on the microarray for every 17 mer present on the tester DNA. Amplified fluorescent tester DNA will hybridize best to its exact complement. See figure below.

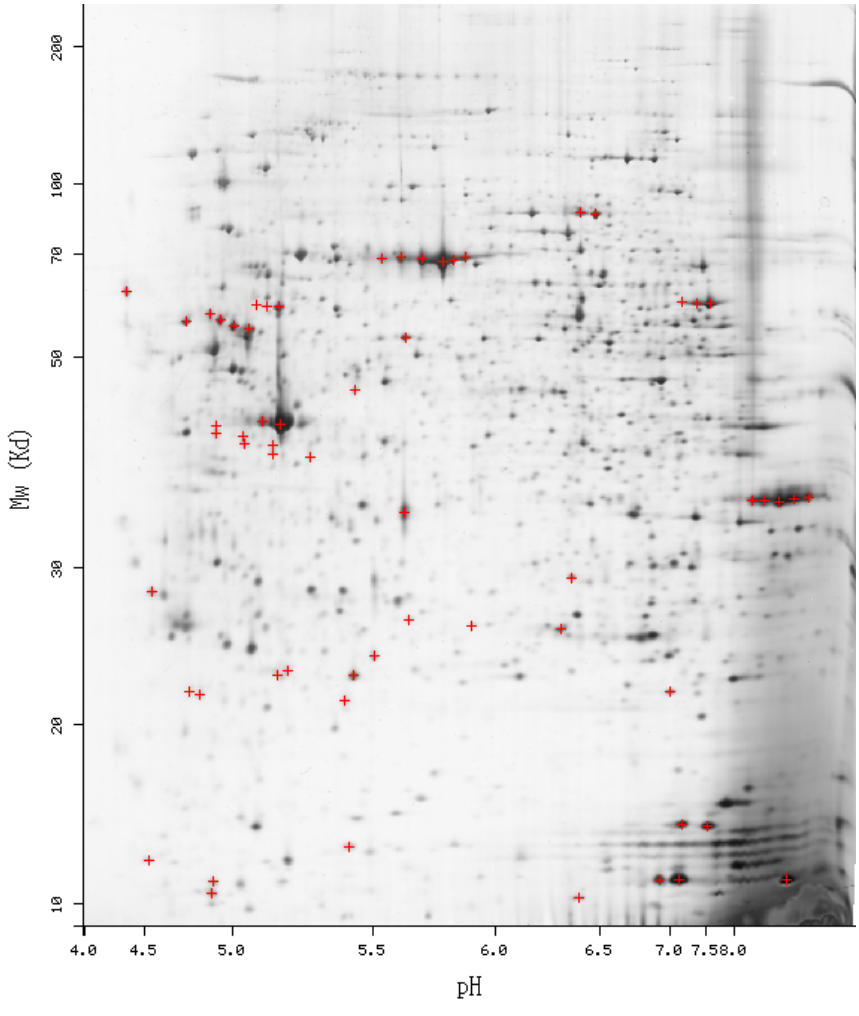
K. SNP detection for linkage analysis. When commonly occuring SNPs are identified in a specific human subpopulation, microarrays are used to screen individuals for SNPs that co-segregate with disease (linkage analyses). Diseased and unaffected individuals are analyzed to determine if certain SNPs are more frequently found in diseased individuals (i.e., certain SNPs are close enough to the disease gene so that the disease gene and the SNP are more likely to stay together than to be separated by crossing over during meiosis). When linkage of a SNP with a disease is determined with statistical certainty, then the next step is to identify the actual nucleotide change

responsible for the disease. Genes close by to the SNP are candidates for the disease gene. These candidate genes are then studied to identify those that are actually involved in the pathogenesis of the disease. The nucleotide sequence of the disease gene can be used as a risk predictor or in some cases as an actual diagnostic indicators. In some cases, the protein products of the genes may be good molecular targets for therapeutics.

L. Microarrays for diagnosis of disease – A cystic fibrosis DNA chip, containing hundreds or thousands of immobilized probes with all known mutations for the cystic fibrosis gene is used to screen cystic fibrosis patients or carriers for specific types of mutations. CF is a common recessive genetic disease due to nucleotide variations present in a single gene. The presence of two CF alleles is diagnostic of the disease. In contrast, the majority of "disease genes" indicate only a risk estimate or a probability that disease will occur; other determinants of disease include environmental factors and the genetic background against which the variant gene under consideration is found (i.e., most diseases are determined by multiple genes and environmental factors). In the next several years, increasing numbers of DNA chips are expected to be available for diagnosis of different diseases from viral or bacterial infection to genetic diseases and cancers.

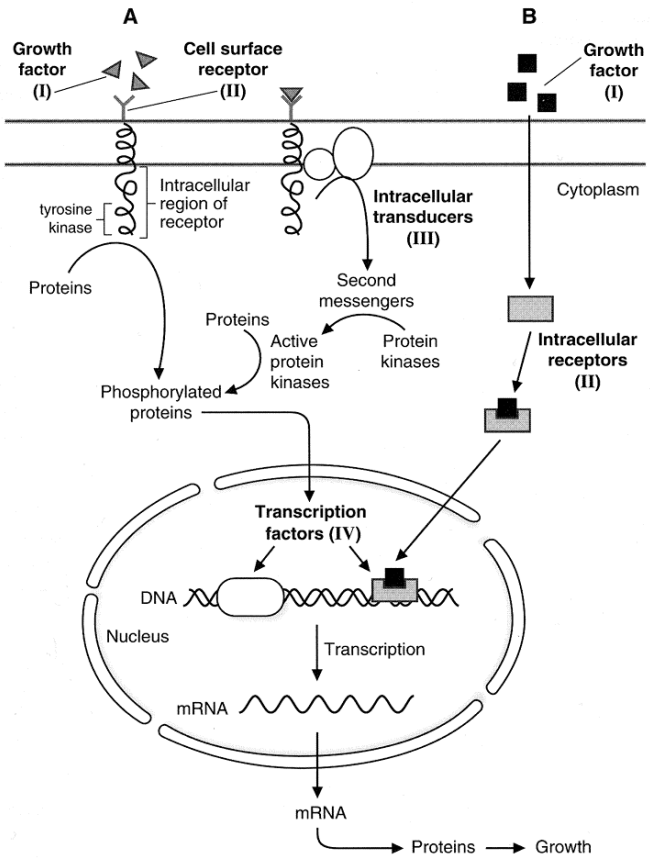
III. **The proteome (the repertoire of proteins produced by a cell) is also being compared among tissues and between normal and diseased tissue.** Amino acid sequence of proteins identified in such screens are used to search genomic sequence databases in order to determine what is known about the protein or its relatives.

Example: a small peptide sequence was obtained from a protein that was shown to be present in cerebral spinal fluid (CSF) and correlated with Alzheimer's disease progression. This protein was identified by comparing the protein profile of Alzheimer's patient's CSF with normal individuals by two-dimensional electrophoresis (see adjacent gel). Spots that differed between the samples were enzymatically chopped into pieces and subjected to mass spectrometry techniques that determine the peptide sequence. From the peptide sequence, bioinformatics, PCR and cloning are being used for further studies of the gene, mRNA and protein.



IV. Bioinformatics – The application of computer technology for the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information. Bioinformatics melds molecular biology with information technology and enables the use of genomic information for understanding human diseases. Improved understanding of human disease leads to new molecular targets for drug discovery and diagnostics development.

A. What can bioinformatics do? Vast amounts of nucleotide sequence of genomic DNA and cDNA is available in databases. Increasing amounts of DNA sequences are being produced daily. Genome sequencing projects that are completed or are underway include the genetic model systems for man such as yeasts, worm, fly, zebra fish and mouse, as well as mosquito, livestock, plants, and infectious organisms. In the model systems, genetic experimentation has provided, and is providing, vast amounts of information on the role of genes in development, adult function and disease. [The amount of similarity of amino acid sequence between man and lower organisms is striking.] This information is giving us a clear picture, gene by gene, of cell and tissue functions. A complete understanding of the function of each protein will enable unimaginable therapeutic options. How is bioinformatics helping us get there? Computational algorithms process genomic sequence to identify genes and predict their intron/exon structure, mRNA, and protein product. Many of the databases are publicly available and contain, for each gene, information that includes:



1. Gene location within the genome.
2. Intron/exon structure.
3. Predicted mRNA.
4. Predicted amino acid sequence.
5. Biological information known about the protein:
   a) Interactions with other proteins.
   b) Position in signaling pathways.
   c) Tissue distribution of expression.
   d) Effect of gene knock-out or specific mutations in model organisms.
   e) Three dimensional structure of the protein. The three-dimensional structure of proteins is perhaps the slowest part of full elucidation. It is slow because of the protein's need to be purified in large quantities and crystallized, and crystallization is often very difficult to obtain. Then the crystal is bombarded with X-rays to obtain a

diffraction pattern that is analyzed to produce a three-dimensional structure. Although this is laborious, there are a initiatives to speed up the process. The three-dimensional structure is important because it can be used in computer modeling studies to identify small chemical compounds that have a high probability to directly interact with receptors or enzymes (i.e., may be good drugs).

6. Related genes in man and other species and related biological information.
7. Polymorphisms and their association with disease.
8. The bioinformatic databases are interrelated and will be updated indefinitely. Thus, in the foreseeable future, the combination of molecular biology and bioinformatics will provide a detailed picture of the functioning of normal cells and tissues, as well as the molecular causes and molecular sequela of diseases. Knowing the roles of proteins (~30,000), how genetic variants of specific proteins predispose to disease, and the molecular response of cells and tissues to disease insults, will allow determinations of individuals' predisposition to diseases. This knowledge will lead to new diagnostics and therapeutics for specific diseases.

V. Identification of proteins that are candidates for drug targets.

A. Drug discovery, a historical perspective.

1. Started from a biochemical pathway implicated in a pathophysiological process.
2. Characterized rate–limiting step in the pathway.
3. Purified enzyme from animal tissues.
4. Screened collections of structurally diverse small molecules to find enzyme inhibitors.
5. Optimized lead compounds by medicinal chemistry.
   a) bioavailability
   b) target specificity for the target enzyme.
6. Drugs that act at receptors were identified by a similar process.

B. Drug discovery and development – present and future perspective.

1. Majority of human genes now available as targets. The challenge now is to identify the best potential targets (enzymes and receptors); those that when the drug (ligand) is bound are likely to produce cures or reduce symptoms. First, the genes that cause disease and the genes involved in the symptoms of disease will be identified:

a) Animal model systems.

b) Genome scans in linkage studies.

c) Gene expression profiling, i.e., which genes are up– or downregulated in disease vs. normal tissue. Expected to reveal many drug targets. Highly selective gene expression, as well as sequence homology to a known gene family, can provide a convenient shortcut for implicating a target in a given pathway or disease.

(a) cause of the pathophysiology (targeting these may modify the disease).

(b) result of the disease (targeting of these may alleviate symptoms).

VI. Sources:

Nature, 2000 volume 21 supplement pp 1 – 60.

Going global p 1 B Phimister

Microarrays and macroconsequences p 2 F S Collins

Array of hope pp 3 - 4 E S Lander

Molecular interactions on microarrays pp 5 - 9 E Southern, K Mir & M Shchepinov

Expression profiling using cDNA microarrays pp 10 - 14 D J Duggan, M Bittner, Y Chen, P Meltzer & J M Trent

Making and reading microarrays pp 15 - 19 V G Cheung, M Morley, F Aguilar, A Massimi, R Kucherlapati & G Childs

High density synthetic oligonucleotide arrays pp 20 - 24 R J Lipshutz, S P A Fodor, T R Gingeras & D J Lockhart

Options available — from start to finish — for obtaining expression data by microarray pp 25 - 32 D D L Bowtell

Exploring the new world of the genome with DNA microarrays pp 33 - 37 P O Brown & D Botstein

The genetics of cancer—a 3D model pp 38 - 41 K A Cole, D B Krizman & M R Emmert–Buck

Resequencing and mutational analysis using oligonucleotide microarrays pp 42 - 47 J G Hacia

DNA microarrays in drug discovery and development pp 48 - 50 C Debouck & P N Goodfellow

Gene expression informatics —it's all in your mine pp 51 - 55 D E Bassett, M B Eisen & M S Boguski

Population genetics—making sense out of sequence pp 56 - 60 A Chakravarti

Websites of the National Cancer Institute.